#### **Lawrence Livermore National Laboratory**

## Lustre at Scale The LLNL Way

D. Marc Stearman



Lustre Administration Lead Livermore Computing - LLNL

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-PRES-403067

# **Topics**

- Project Structure
- LLNL Computing Platforms
- Network Design and Topology
- Software Release Methodology
- Operation and Management
- Next Steps
- Hyperion
- Concerns





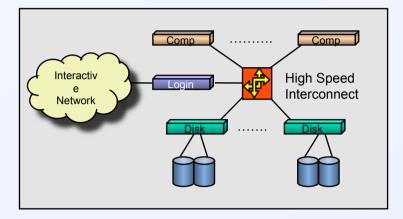
#### **Project Structure -** "Who are all these people, and what are they doing here?"

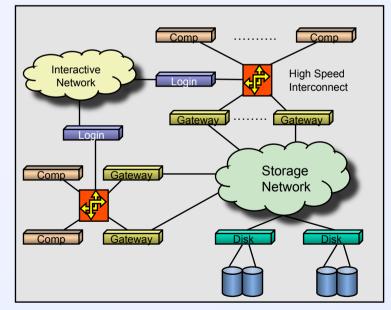
- Project Lead Mark Gary
- Administration/Operations Team Marc Stearman
  - Responsible for daily system administration, cluster integration, upgrades, hardware repair, user application support
  - 4 Full Time Employees + HW Repair Team
- Software Development Jim Garlick
  - Responsible for bug fixing, build, QA, and tool development
  - 4 Full Time Employees



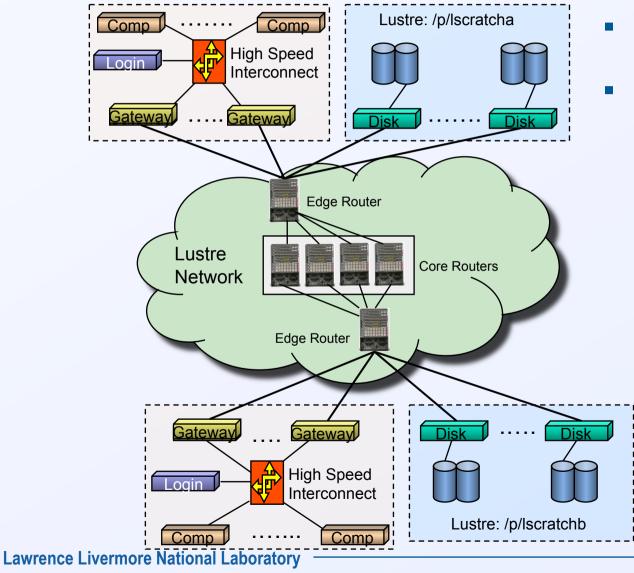
# **HPC Compute/Storage Pairing Philosophies**

- Islands of Storage
  - Many HPC clusters, with dedicated attached storage
  - Data access internal to cluster
  - High performance Uses local high-speed interconnect
  - Many copies of data running on multiple clusters
- Peninsulas of Computation
  - Many HPC clusters using shared network storage
  - User convenience Fewer copies of data
  - Redundant if one file system is down, others are still available
  - Extra network latency



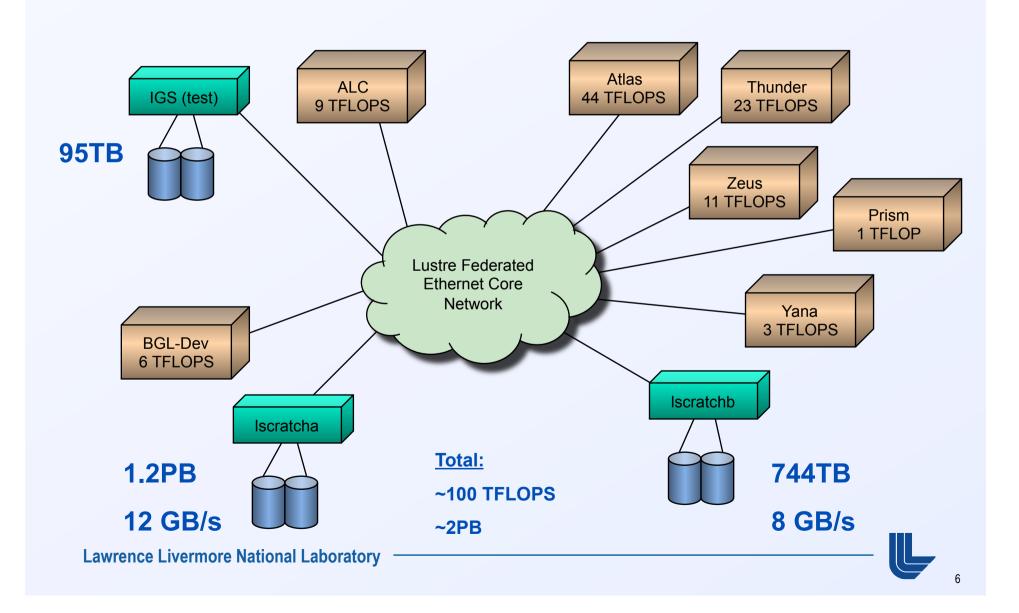


### **Peninsulas Explained**

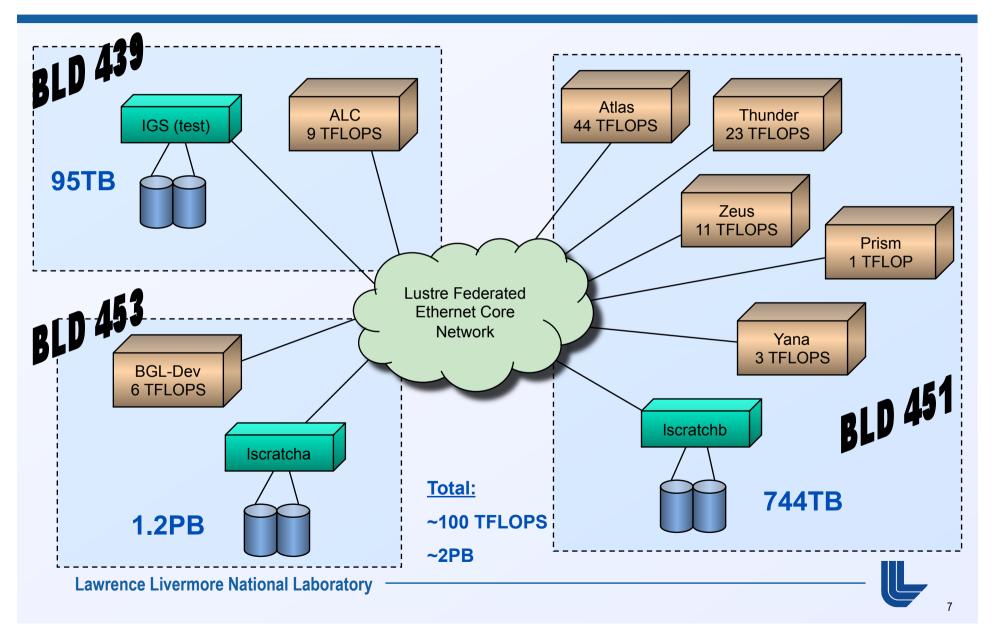


- All storage visible via storage network
- Compute clusters
  have some Lustre
  file systems more
  "local" than others

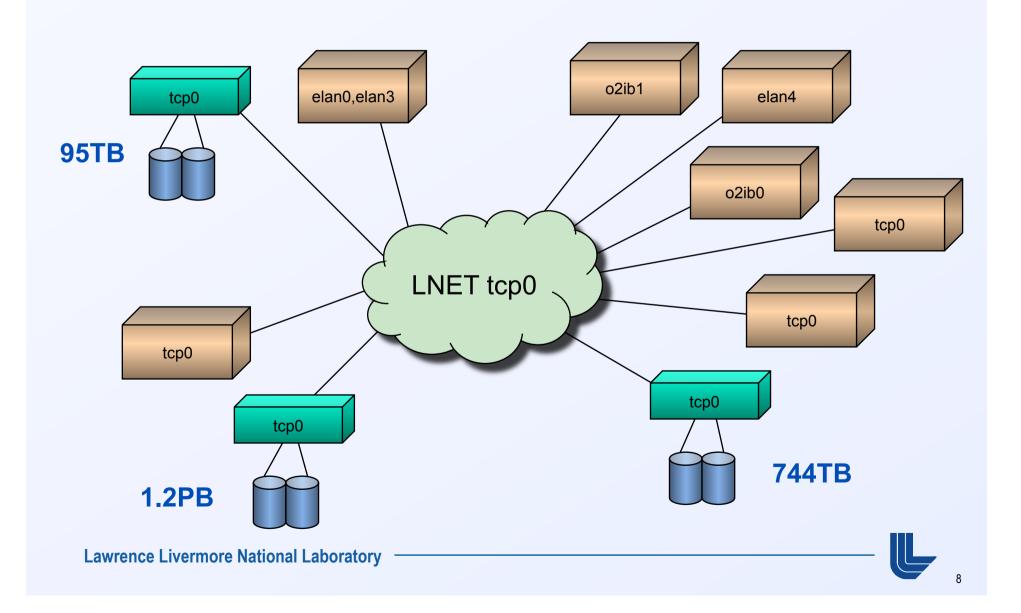
### **Livermore Computing - Open Computing Facility**



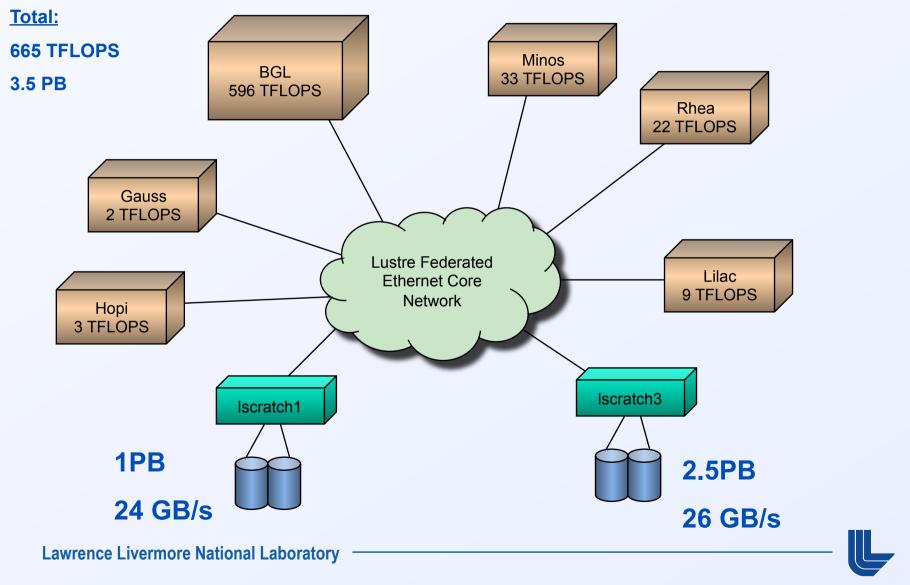
#### **Livermore Computing - Open Computing Facility**



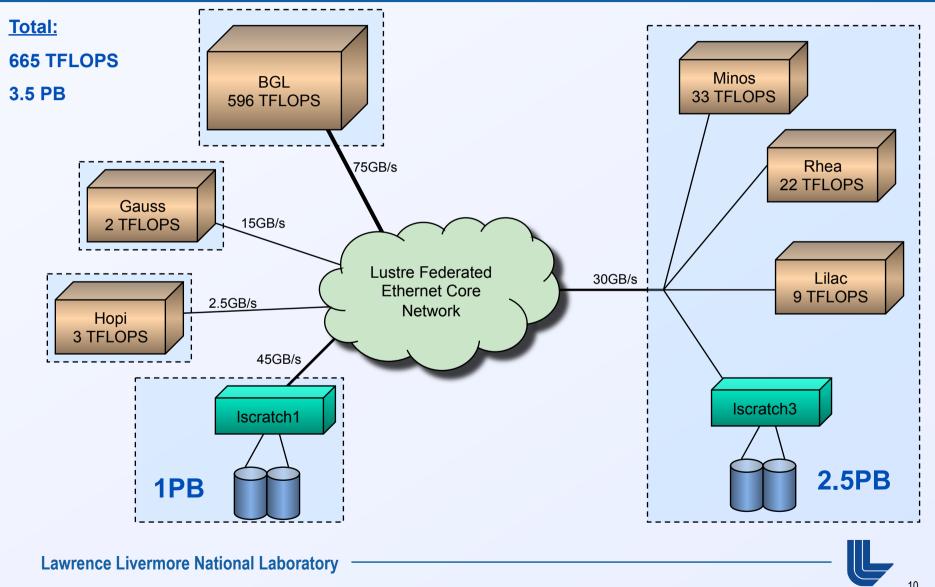
### LNET view of the world - Open Computing Facility



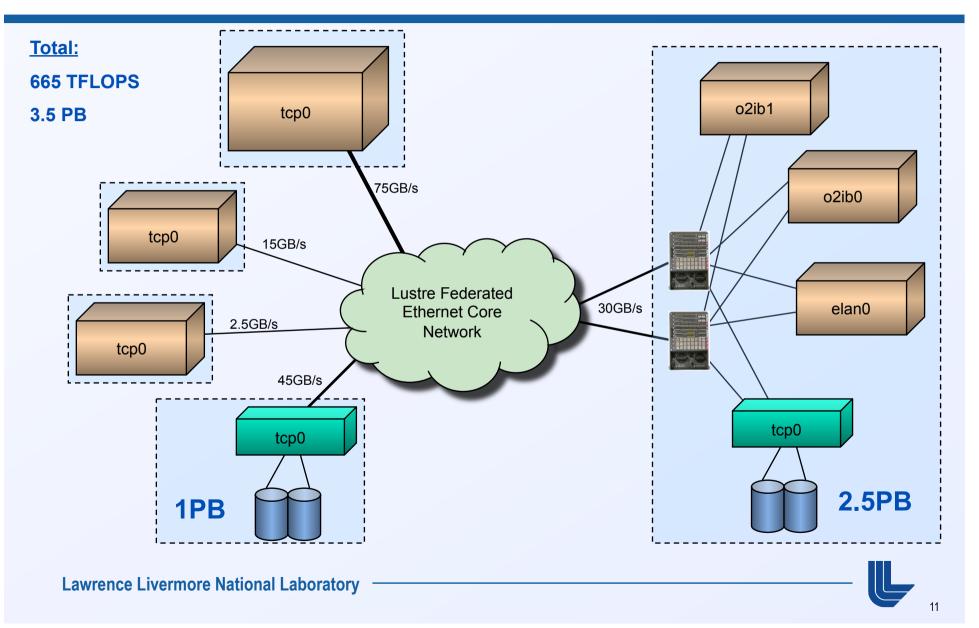
### **Livermore Computing - Secure Computing Facility**



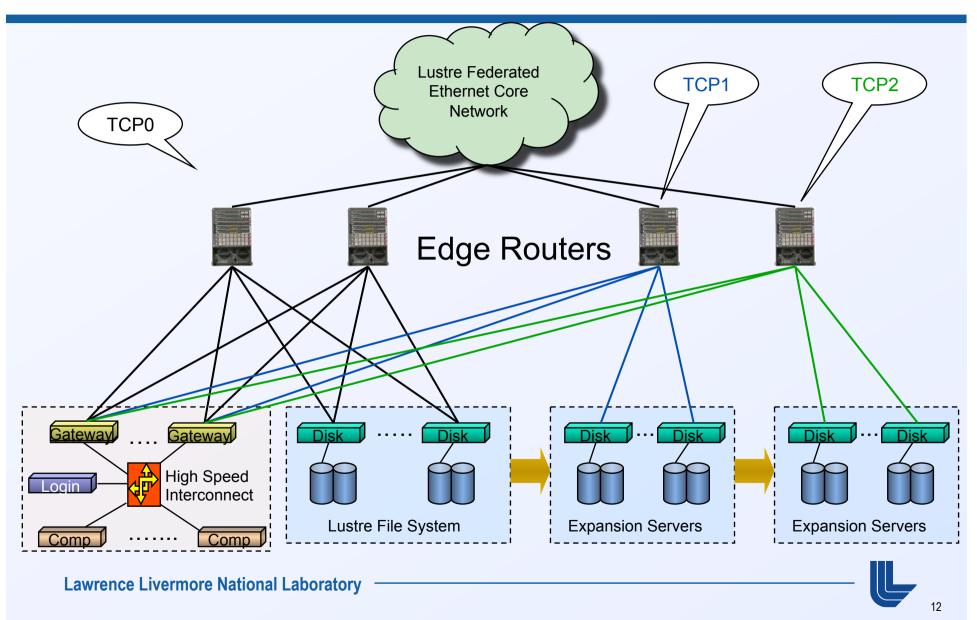
#### **Livermore Computing - Secure Computing Facility**



#### **LNET view of the world - Secure Computing Facility**



#### Livermore Computing - Next Steps (LNET)

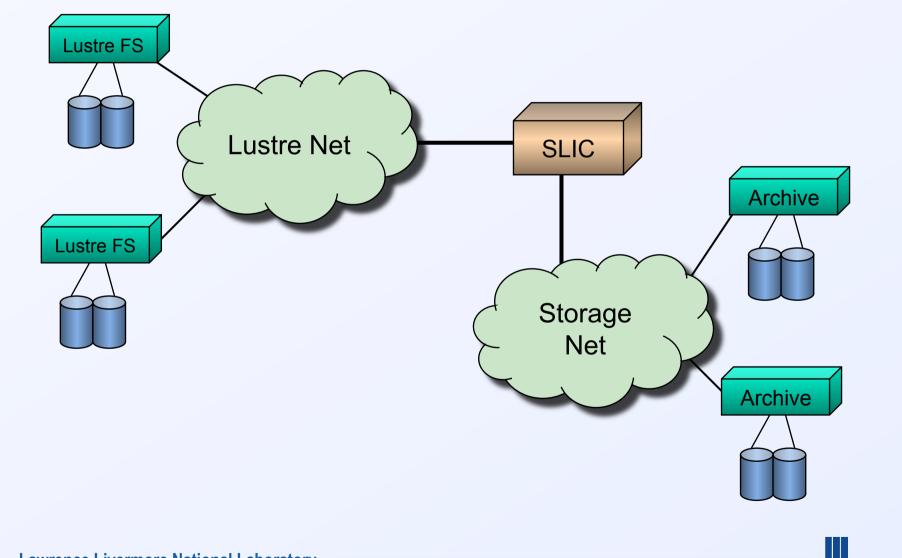


# Livermore Computing - Next Steps (LNET)

- Benefits
  - Reduced network traffic through the core switches
  - Keeps traffic local to switch backplanes and reducess latency
  - Saves Money!
- Issues
  - Need dynamic NIDs
    - Routes can be added dynamically via 'lctl' on cluster with routers, but tcp-only clients on tcp0 need unload of lustre modules to add tcp1 and tcp2 NIDs
    - With multiple lustre filesystems mounted, ALL filesystems must be unmounted so all jobs are killed limiting production work
  - Makes LNET configuration more complex (But we already have that)



#### **SLIC - Storage Lustre Interface Cluster**



## **Software Release - Testing Methodology**

- Sanity test in Build Farm
- Small scale testing in Testbed
- Mid scale testing on BGL-Dev, and new 64 node cluster
- Large scale testing on ALC (~400 clients)
- Giant scale testing on Atlas (1100 clients) during DST (Dedicated System Time)
- Wide variety of tests IOR, iozone, fsx, MIB, mdtest, simul, PIOS, various reproducers
- Developers are using Lustre for their /home file system
- We carry 150+ patches from base releases to make lustre work on our production systems - This requires a great deal of testing resources

## **Software Release - Rollout Methodology**

- Midway through migration from 1.4 to 1.6
- Servers before Clients
  - All servers now running 1.6.2-30chaos (1.6.2 + 160 patches)
  - All x86\_64 clients and BGL moving to 1.6.2-30chaos
  - All i686 and ia64 clients will remain at 1.4.8-22.3chaos (1.4.8 + 156 patches) until they retire (about 6-9 months from now)
- The 1.4 clients can mount the 1.6 file systems because they were created under 1.4, and migrated to 1.6
- If we field a new file system we have three options for our legacy clients
  - Create it at 1.4 and migrate to 1.6
  - Develop a tool that will read the 1.6 on-disk configuration, and write out 1.4 style config
  - Not mount legacy clients



### **Operations and Management - "Sameness"**

- Storage Scalable Units (SSUs)
  - Concept borrowed from Compute Clusters
  - Building blocks used to build/expand file systems
  - Makes deployment smooth and quick
- Hardware Repair team



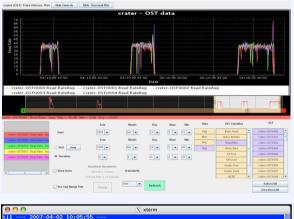
- Operations Personnel handle all HW repair actions with minimal SysAdmin intervention
- Operator training
  - Training courses developed locally
  - Knowledge Base on local wiki
  - Goal to reduce off-hours pages
  - Numerous scripts for testing and problem determination
- LMT v2

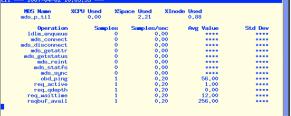
## LMT v2 - http://sourceforge.net/projects/lmt/

File Configure

Start with xwatch-lustre functionality, then add:

- New views (OSS, Filesystem, Router Group, ...)
- Plotting capability (historical trends, heart-beat, ...)
- Customization features
- Full-system health "at a glance"





Lawrence Livermore National Laboratory

mds	p_lscratch-	1 2008-	04-18 0	4:17:45.0		S	OST	2008-04	-10 04:1	7:45.0		View		
PU	SKB			%Inodes		Ost Name	Read	Write	NEPU	2688	hinodes	These		
5.35	3.21			35.89		OST Jevi48 3	Rate 0.00	Rate 0.00	****	87.77	2.86	Router Gro	up 2-2	008-04-1
eration	Samples	Sample	Avg	Std Dev	Units	OST_levi45_4	0.00	0.00	****	87.17	7.78	Name	Rate	NCPU
2021/2021		/Sec	Value			OST_levi46_4	0.08	8.60	****	86.71	8,41	atlas64	0.04	0.10
m_bl_caliback	0	0.00	0.00	0.00	usec	OST_Jevi47_4	0.08	0.00		07.51	8.30	atlas65	0.04	8.12
m_cancel	0	0.00	0.00	0.00	usec	OST_levi48_4	0.00	0.00		87.65	7.79	atlas66	0.04	.0.12
m_convert	0	0.00	0.08	0.08	usec	OST_levi45_5	0.00	8.00		82.75	6.98	atlas67	0.04	0.15
m_cp_caliback	0	0.00	0.08	0.08	usec	OST_levi46_5	0.00	8.80	-	87.85	7.53	atlas68		0.12
m_enqueue	29274	5,862.40	692.86	0.08	usec	OST_levi47_5	0.00	0.00		87.66	8.14	atlas59	0.05	0.22
m_gLcallback	0	0.00	0.08	0.00	usec	OST_levi48_5	0.00	0.01		87.11	8.32	atlas70	0.05	0.17
g_origin_handle_cancel	0	0.00	0.00	0.00	usec	OST_Jevi45_6	0.00	8.80		82.58	7.00	atlas71	0.04	0.20
s_close	0	0.00	0.00	0.00	usec	OST_levi46_6	0.00	8.80		86.95	8.32	atlas504	0.05	8.12
s_connect	0	0.00	0.00	0.00	usec	OST_levi47_6	0.00	8,80		17.45	8.45	atlas505	0.04	0.12
is_disconnect	0	0.00	0.00	0.08	usec	OST_levi4.8_6	0.00	8.80		81.95	7.00	atlas506	0.04	0.10
ts_dose_writing	0	0.00	0.08	0.08	usec	OST_levi45.7	0.00	0.01		86.71	8.42	atlas507	0.04	0.15
da_getattr	0	0.00	0.08	0.08	usec	OST_levi46_7	0.00	0.00	ana	87.55	8.23	atlas508	0.05	0.17
\$3_getattr_lock	0	0.00	0.08	0.08	usec	OST_levi47_7	0.00	0.00		87.63	7.77	atlas503	0.05	0.17
ds_getstatus	0	0.00	0.00	0.08	usec	OST_levi48_7	0.00	8.00	anter .	87.05	7.81	atlas510	0.04	0.12
ts_getcattr	0	0.00	0.00	0.00	usec	OST_levi45_8	0.00	8.00		38.84	7.69	atlas511	0.05	0.15
ds_pin	0	0.00	0.08	0.08	usec :	OST_levi46_8	0.00	8.80		87.75	7.86	atlas640	0.04	0.12
ds_quitacheck	Û	0.00	0.00	0.00	usec	OST_levi47_8	0.12	3.74		17.01	7.76	atlas641	0.05	0.17
és_quitarti	0	0.00	0.08	0.08	usec	OST_levi48_8	0.00	8.80		32.24	7.11	Router Gro	un d 3	008-04-1
ds_readpage	.0	0.00	0.00	0.00	usec	05T_levi45_9	0.00	0.00	-	87.62	7.76	Name	Rate	SEPU
ds_reint	10	3.20	1,045.98		usec	OST_levi46_9	0.00	0.00	-	87.55	8.45	78110	46.93	2.92
ds_set_info	0	0.00	0.00	0.08	usec	OST_Jevi47_9	0.00	8.80	-	86.89	8.32	zeutl	52.22	8,77
ds_setxattr	0	0.00	0.08	0.08	usec	OST_Invi48_9	0.00	0.00		87.31	8.33	zeut2	54.54	2.07
ds_statts	0	0.00	0.08	0.08	usec	OST_levi45_10		0.00		38.04	8.28	zeus3	41.57	0.62
ds_sync	0	0.00	0.08	0.08	usec	OST_levi46_10		0.00	****	87.28	7.03	zeus4	48.99	0.82
ds_ungin	0	0.00	0.08	0.08	usec.	OST_levi47_10		1.00		17.31	7.81	2005	1.71	1.17
id_ping	100	28.00	47.44	22.29	usec	OST_levi48_10		0.00		12.44	7.18	2eus6	57.31	0.90
rt_clase	0	0.00	0.08	0.08	usec	05T_levi45_11		0.10	-	19.61	8.32	zeui7	59.52	0.95
t_consect	0	0.00	0.00	0.00	usec	051_levi45_11 051_levi46_11		0.00		86.54	8.32	AGGREGATE	362.83	44444
rt_create	0	0.00	0.08	0.08	usec			0.00		87.12	8.29	MASIMUM	59.52	7.97
st_destroy	0	0.00	0.08	0.08	usec	OST_levi47_11 OST_levi48_11		0.00		87.37	8.47	MINIMUM	1.71	0.62
st_disconnect	0	0.00	0.08	0.08	usec			0.00		86.95	8.48	AVERAGE	45.35	1.91
t_get_isfo	0	0.00	0.08	0.08	usec	OST_levi45_12						AVENAGE	92.53	1.21
t_getattr	0	0.00	0.00	0.00	usec	OST_levi46_12		8.80		87.35	8.39			
t_open	Û	0.00	0.00	0.00	usec	OST_levi47_12		0.00	****	36.64	0.45			
t_punch	0	0.00	0.00	0.08	usec	OST_levi48_12		0.00		17.22	8.36			
t_gustacheck:	0	0,00	0.08	0.08	usec	AGGREGATE	0.37	0.09			100 C 100 C 100 C			
rt_quetacti	0	0.00	0.00	0.00	usec	MAXIMUM	0.25	3.89	-	89.68	8.75			
st_read	0	0.00	0.08	0.08	usec	MINIMUM	0.03	8.80		80.12	6.07			

Router Name	BN MB/s	XCPU Used
adev4	38.62	11.60
adev5	42.32	12,10
adev6		
Naximum	42.32	12,10
Average	26,98	7,90
Aggregate	80,95	
Router Name	BN MB/a	XCPU Used
odev8		
odev9		
Neximu		
Average	0.00	0.00
Aggregate	0.00	
Router Name	BN MB/s	XCPU Used
tdev5	117.87	6,58
tdev6	116,67	6,73
Maximum	117,87	6.73
Average	117.27	6,65
Aggregate	234.54	

T_ilc2				XSpace Used	XInodes Used
I_IICZ	54,25	0.00	7.57	12.29	0.00
T_ilc3	83,60	0.00	14.56	11.69	0.00
T_ilc4	90.03	0.00	14.37	11.51	0.00
T_ilc5	59,60	0.00	8,95	11.16	0.00
laximum	90.03	0.00	14.56	12.29	0.00
verage	71.87	0.00	11.36	11.67	0.00
regate	287.48	0.00			
	T_ilc4 T_ilc5  aximum werage	T_ilc4      90.03        T_ilc5      59.60        axinum      90.03        verage      71.87	T_i1c4 90.03 0.00 T_i1c5 59.60 0.00 aximum 90.03 0.00 werage 71.87 0.00	T_i1c4      90.03      0.00      14.37        T_i1c5      59.60      0.00      8.95        aximum      90.03      0.00      14.56        verage      71.87      0.00      11.36	T_ilc4      90.03      0.00      14.37      11.51        T_ilc5      59.60      0.00      8.95      11.16        aximum      90.03      0.00      14.56      12.29        verage      71.87      0.00      11.36      11.67

## **Livermore Computing - Next Steps**

- Filesystem Requirements
  - Dawn ( .5 PFLOPS): 96 GB/s, ~4PB
  - Sequoia (10 PFLOPS): 512 GB/s, ~50PB
- Router Cluster
  - Sequoia may have an Infiniband network
  - Need to bridge Infiniband with legacy 10GE Net
- Evaluating ZFS (Both User Space and In Kernel)
- Failover





- Will replace ALC as large scale lustre test cluster
- Hyperion is a partnership with the vendor community, including Sun/CFS
- Will test and evaluate emerging technologies OFED, Virtualization, Lustre, QDR Infiniband, 40/100 GE, etc
- 1152 nodes, 9K cores, 120 TFLOPS
- 1.6 PB disk @ >36 GB/s
- Two Lustre Networks
  - Infiniband
  - 10 Gigabit Ethernet



#### Concerns

- Space Management
  - How do you manage a 50 PB file system?
  - Quotas?
  - Purges?
  - Conventional tools do not scale
  - Is, tar, cp, rsync, etc
  - Moving data to Archive efficiently
- ZFS
  - Performance: User space vs. Kernel space
  - Migration from Idiskfs to ZFS
- Scaling concerns
  - Metadata Performance
  - Adaptive Timeouts
  - Multi/Many core parallelization



#### Questions



